С. В. Попов

ЗАКОНОМЕРНОСТЬ РАСПРЕДЕЛЕНИЯ ДОКУМЕНТОВ ПО КЛАССАМ БЛИЗОСТИ К БИНАРНЫМ ВЕКТОРАМ ТЕРМИНОВ ПОИСКОВЫХ ЗАПРОСОВ В ПОЛИТЕМАТИЧЕСКИХ МАССИВАХ КОРОТКИХ ДОКУМЕНТОВ

Одно из важных открытий в области библиометрии и наукометрии было сделано С. Брэдфордом, сформулировавшим закон рассеяния статей по одной тематике в системе периодических изданий [1]. Можно сказать, что эта закономерность характеризует нарастание затрат (количество просматриваемых статей) при попытке найти всю информацию по заданной тематике, начиная поиск с наиболее профильных периодических изданий и переходя к менее профильным.

Настоящее исследование направлено на определение таких затрат при работе с ранжирующими информационно-поисковыми системами (ИПС) с бинарной метрикой (о таких системах можно прочитать, например, в работах [2] и [3]), в которых в качестве документов исходного поискового массива используются короткие тексты (рефераты, новостные сообщения и т.д.) и/или части больших текстов (например, абзацы статей или книг).

В таких ИПС документы исходного массива делятся на классы в зависимости от того, сколько терминов из вектора поискового запроса содержится в каждом из документов. В качестве составляющих вектора запроса и документа выступают ключевые слова, индексы классификаций, дескрипторы, фразы или просто слова естественного языка. При использовании бинарной метрики терминам запроса не приписываются веса значимости, и каждому из них соответствуют два возможных состояния: либо он содержится в данном документе, либо нет [2, с. 143].

Автором настоящего исследования утверждается, что независимо от языка, на котором написаны документы (при условии, что вектор запроса состоит из терминов того же языка), для политематических массивов размеры классов (количество документов в классе) подчиняются следующей закономерности, независимо от тематики поискового запроса:

$$n\ (\mu) = n_0 \times \left(1 - \frac{n_o}{N}\right)^{\mu},$$

где n (μ) − количество документов класса, бинарная мера близости которого к вектору запроса (количество совпавших терминов) равна μ; n_o − количество документов исходного массива, не имеющего ни одного общего термина с вектором запроса; N − общее количество документов исходного массива.

Естественным условием проявления закономерности является то, что термины вектора запроса выбираются неслучайно и отражают лек-

сику исследуемой тематической области. Исключается использование в векторе запроса предлогов, союзов, частиц, но допускается использование терминов, присущих не только исследуемой, но и другим тематическим областям. Окончательное определение условий проявления закономерности требует продолжительных экспериментов, но уже первые опыты показали хорошее совпадение теоретических и экспериментальных результатов (см. раздел «Экспериментальная проверка»).

Теоретическое обоснование

Документальный векторный поиск с бинарной метрикой основывается на определении степени смыслового соответствия между содержанием документа и информационным запросом, которые выражаются следующим образом:

$$\mu = 0;\ 1;\ 2;\ 3;\ \ldots\ldots\ k, \tag{1}$$

где k – максимальное количество (без учета повторений) терминов вектора запроса, присутствующих в документах (документе) исходного поискового массива.

Распределение документов исходного поискового массива в соответствии со значениями μ обозначим через p(μ),

$$p(\mu) = \frac{n(\mu)}{N} \tag{2}$$

где n(μ) – количество документов исходного массива, содержащих ровно μ терминов из вектора запроса; N – общее количество документов исходного массива.

По определению,

$$\sum_{\mu} n(\mu) = N \tag{3}$$

Следовательно,

$$\sum_{\mu} p(\mu) = 1 \tag{4}$$

При поиске в различных политематических массивах документов по любому тематически избирательному запросу математическое ожидание величины степени смыслового соответствия документа и запроса E(μ) (конечно, при использовании неслучайных и осмысленных терминов в запросе) близко к нулю. Где

$$E(\mu) = \sum_{\mu} \mu \times p(\mu)$$

Значение E(μ) достигает минимума при самом избирательном запросе, когда вектор запроса характеризует только один термин и на этот запрос выдаётся ровно один документ, при этом значение E(μ) =

1/N. Значение E(μ) достигает максимума при самом неизбирательном запросе, когда вектор запроса характеризуется всеми словами, используемыми во всех документах поискового массива, при этом значение E(μ) = M, где M – среднее количество слов в документах поискового массива. Например, если все документы поискового массива содержат по 400 слов, то E(μ) = 400. Мы предполагаем, что при использовании неслучайных и осмысленных терминов в запросах, каждый из которых характеризует конкретную тематическую область, при поиске в политематических массивах значения E(μ) будут колебаться вокруг некоего среднего значения $E_{ср}$ близкого к нулю. Например, в проведенных экспериментах (см. далее) значение $E_{ср}$ = 0,25. Требуется проведение значительного количества дополнительных экспериментов для уточнения величины $E_{ср}$.

Итак мы предполагаем, что векторный поиск в политематических массивах документов можно охарактеризовать условием:

$$\sum_{\mu} \mu \times p(\mu) = E_{cp} = const \qquad (5)$$

Во многих случаях вместо величин p(μ) удобнее и нагляднее использовать величины:

$$I(\mu) = -\log p(\mu), \qquad (6)$$

названные в теории информации ***количеством информации случайной величины*** $\boldsymbol{\mu}$ [4, с. 24].

Математическое ожидание:

$$S(\mu) = -\sum_{\mu} p(\mu) \times \log p(\mu) \qquad (7)$$

называется ***энтропией распределения*** p(μ) [4, с. 25].

Можно найти наиболее вероятное распределение $\overline{p}$(μ) для всех векторов осмысленных тематических запросов, используя метод поиска наиболее вероятных распределений, разработанный в статистической физике [5, с. 320–327]. Суть метода заключается в поиске максимума энтропии (7) при заданных условиях. В нашем случае такими условиями являются условия (4) и (5). Основание логарифма в выражении энтропии для наших задач несущественно, далее будем использовать натуральные логарифмы. Применим метод Лагранжа для поиска максимума функции:

$$S(\mu) = -\sum_{\mu=0}^{k} p(\mu) \times \ln p(\mu)$$

Рассмотрим S(μ) как функцию от (k+1)-ой независимой переменной p(μ) и найдем ее максимум при условиях:

$$\begin{cases} \sum\limits_{\mu=0}^{k} p(\mu) = 1 \\ \sum\limits_{\mu=0}^{k} \mu \times p(\mu) = const \end{cases} \qquad (8)$$

В данном случае лагранжиан имеет вид:

$$L = -\sum_{\mu=0}^{k} p(\mu) \times \ln p(\mu) + \lambda \times \left(1 - \sum_{\mu=0}^{k} p(\mu)\right) + \alpha \times \left(const - \sum_{\mu=0}^{k} \mu \times p(\mu)\right),$$

где λ и α – множители Лагранжа.

Дифференцируя L частным образом по каждой из переменных p(μ), получаем следующую систему из (k+1)-го уравнения:

$$\ln\ p(\mu) + 1 + \lambda + \mu\ \alpha = 0\ ,\ \mu = 0 \ldots k$$

Откуда

$$\overline{p}(\mu) = \beta\ e^{-\alpha\mu} \qquad (9)$$

где $\beta = e^{-(1+\lambda)}$. Множитель β можно определить по формуле (9), положив μ = 0, тогда для всех распределений p(μ)

$$\beta = p(0) = p_0, \qquad (10)$$

тогда

$$p(\mu) = p_0 \times e^{-\alpha\mu}$$

Множитель α определим из условия нормировки (4):

$$\sum_{\mu} p(\mu) = \sum_{\mu} p_0 \times e^{-\alpha\mu} = 1,$$

или, учитывая (1),

$$p_0 \times (1 + e^{-\alpha} + e^{-2\alpha} + \ldots + e^{-k\alpha}) = 1$$

Сумма в скобках равна сумме геометрической прогрессии со знаменателем $e^{-\alpha} < 1$, и условие нормировки преобразуется к виду

$$\frac{p_0 \times (1 - e^{-\alpha k})}{1 - e^{-\alpha}} = 1$$

Предположим, учитывая свойства экспоненты с отрицательной степенью, что значение $e^{-\alpha k}$ пренебрежительно мало. В дальнейшем мы подтвердим корректность такого предположения.

Тогда условие нормировки преобразуется к виду

$$\frac{p_0}{1-e^{-\alpha}}=1$$

и

$$\alpha = -\ln\left(1 - p_0\right) \qquad (11)$$

Экспериментальные исследования, представленные ниже, показывают, что даже для минимального значения k = 7 величина – αk = – 1,6×7 = – 11,2, а $e^{-\alpha k} \approx 0{,}000014$.

Учитывая (10) и (11), получим из (9) окончательное выражение для наиболее вероятного распределения

$$\bar{p}\left(\mu\right)= p_0 e^{\ln\left(1-p_0\right)\times\mu} = p_0\left(1-p_0\right)^{\mu} \qquad (12)$$

Или

$$n\left(\mu\right)= n_0 \times\left(1-\frac{n_o}{N}\right)^{\mu}, \qquad (13)$$

где n(μ) – количество документов класса, бинарная мера близости которого к вектору запроса равна μ; n_o – количество документов исходного массива, не имеющего ни одного общего термина с вектором запроса; N – общее количество документов исходного массива.

Экспериментальная проверка

В таблице 1 приводятся экспериментальные и теоретические значения p(μ) для различных тематических запросов. Эксперименты проводились на ИПС «SOVA+» [6], разработанной в ФГУП ИПР «Информэлектро». Экспериментальные распределения p(μ) получены при поиске в массиве документов из бюллетеня «Промышленность: 100 новостей», выпускаемого тем же институтом. Размер исходного массива N=4703 документа. Теоретические распределения построены по формуле (12).

Выявленная закономерность дает возможность усовершенствовать разработанный автором метод поиска информации [7].

В таблице 2 приводится список слов, использованных при поиске по теме «Экология».

В таблице 3 приводится фрагмент бинарной матрицы «запрос-документ» для списка слов по теме «Экология».

Таблица 1

Результаты сравнения экспериментальных и теоретических значений p(μ) для различных тематических запросов

Тематика запроса	**Способ получения p(μ)**	**экспериментальные и теоретические значения p(μ) величины μ**									
		0	1	2	3	4	5	6	7	8	9
Экология (49 терминов)	эксп.	0.80	0.16	0.03	0.01	0.00	0.00	0.00	0.00	–	–
	теор.	0.80	0.16	0.03	0.01	0.00	0.00	0.00	–	–	–
Телекоммуникации (33 термина)	эксп.	0.90	0.08	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	теор.	0.90	0.09	0.01	0.00	0.00	–	–	–	–	–
Очистка воды (16 терминов)	эксп.	0.97	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	теор.	0.97	0.03	0.00	–	–	–	–	–	–	–
Металлургия (43 термина)	эксп.	0.59	0.22	0.10	0.05	0.02	0.01	0.01	0.00	0.00	0.00
	теор.	0.59	0.24	0.10	0.04	0.02	0.01	0.00	0.00	0.00	0.00
Информационные технологии (26 терминов)	эксп.	0.88	0.11	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	теор.	0.88	0.11	0.01	0.00	0.00	–	–	–	–	–
Фундаментальная наука (17 терминов)	эксп.	0.73	0.20	0.05	0.01	0.00	0.00	0.00	0.00	0.00	–
	теор.	0,73	0,20	0,05	0,01	0,00	0,00	0,00	0,00	–	–

Таблица 2

Список слов, использованных при поиске по теме «Экология»

№	**Слово**	**№**	**Слово**
1	экологически	26	воздуха
2	экологической	27	вредных
3	отходов	28	мусора
4	технология	29	новая
5	чистых	30	нового
6	серы	31	новую
7	технологии	32	новых
8	безопасности	33	окружающей
9	выбросы	34	отходы
10	обработки	35	охране
11	очистки	36	очистных
12	переработке	37	природопользованию
13	переработки	38	радиоактивных
14	содержанием	39	среды
15	технологию	40	территории
16	установка	41	технологий
17	химических	42	установку
18	чистого	43	химии
19	чистой	44	химического
20	чистый	45	чистая
21	экономические	46	чистые
22	экологический	47	эко
23	экологическую	48	экологических
24	атмосферу	49	экология
25	воды		

Таблица 3

Фрагмент бинарной матрицы «запрос-документ» для списка слов по теме «Экология»

№ п/п док	Номера слов (по табл. 2)																																																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	
1																								1																										1
2		1																																																2
3		1																																																3
4		1																																																4
5		1																																																5
6		1																																																6
7		1																																																7
8		1																																																8
9		1																																																9
10		1																																																10
11		1																																																11
12		1																																																12
13																									1																									13
14																									1																									14
15																									1																									15
16																									1																									16
17																									1																									17
18																									1																									18
19																									1																									19
20																									1																									20
21																									1																									21
22																									1																									22
23																									1																									23
24																									1																									24
25																																																		25
26																																																		26
27																																																		27
28																																																		28
29																											1																							29
30																								1			1																							30
31																								1			1																							31
32				1																									1																					32
33				1																									1																					33
34				1																									1																					34
35										1							1																																	35
36											1					1																																		36
37									1															1			1																							37

Таблица 3 (продолжение)

№ п/п док	Номера слов (по табл. 2)																																																	
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	
38				1													1																	1																38
39				1														1											1																					39
40				1															1															1																40
41				1			1																						1																					41
42				1			1																						1																					42
43									1				1																			1				1														43
44			1									1																1																						44
45							1																									1									1		1							45
46			1									1																					1						1											46
47				1					1															1		1			1																					47
48	1									1																																			1					48
49	1					1								1						1					1	1																								49
50			1				1								1		1										1																							50
51			1	1			1								1																1																			51
52			1									1																					1		1				1											52
53															1		1																1	1					1											53
54		1	1										1																					1		1														54
55			1	1									1				1																							1										55
56	1		1	1	1										1																																			56
57	1					1							1	1				1																																57
58	1			1			1												1																									1						58
59		1															1								1																1						1			59
60	1					1								1						1						1																								60
61		1						1																									1		1		1		1											61
62	1		1	1									1															1																		1				62
63				1																		1		1					1	1				1										1						63
64	1			1	1					1																			1														1							64
65	1		1									1				1				1																		1												65
66		1						1			1																						1		1		1		1											66
67																																																		67

Линия «обрыва» таблицы 3

* * *

Установленная закономерность позволяет оценить нарастание затрат (количество просматриваемых документов) при попытке найти всю информацию по определенной тематике, используя ранжирующую ИПС бинарного типа и начиная просмотр с документов, принадлежащих классу, характеризующемуся максимальной степенью близости к запросу ($\mu = k$). Для запроса по теме «Экология» теоретически оцениваемое нарастание затрат (формула (13), $N = 4703$ документов, значение n_o установлено экспериментально и равно 3726 документам) будет происходить следующим образом: $n(5) = 1$ документ, $n(4) = 6$ документов, $n(3) = 30$ документов, $n(2) = 190$ документов, $n(1) = 750$ документов. Таким образом, для просмотра последнего класса документов потребуется затрат в 3 раза больше, чем для просмотра всех предыдущих классов документов вместе взятых. При этом может оказаться, что именно в этом последнем классе в небольшом количестве содержится наиболее интересная для пользователя информация.

Выявленная закономерность дает возможность усовершенствовать разработанный автором метод поиска информации [7] и повысить эффективность поиска релевантных документов, принадлежащих классам, характеризующимся малыми значениями степени близости к запросу.

Литература

1. Bradford S.C. The documentary chaos // Bradford S.C. Documentation. London: Crosby Lockwood, 1948. P. 106–121.
2. Сэлтон Г. Автоматическая обработка, хранение и поиск информации / Пер. под ред. А.И. Китова. М.: Советское радио, 1973.
3. Кириченко К.М., Герасимов М.Б. Обзор методов кластеризации текстовой информации. Материалы Международной конференции ДИАЛОГ–2004. Т. 1: Теоретические проблемы; http://www.dialog-21.ru/archive\ article.asp? param = 6912
4. Колесник В.Д., Полтырев Г.Ш. Курс теории информации. М.: Наука, 1982.
5. Сивухин Д.В. Термодинамика и молекулярная физика. Т. 2. М.: Наука, 1990.
6. Масликов В.В., Попов С.В., Сергеев А.С. Лингводинамическая система поиска «SOVA+». Свидетельство об официальной регистрации программы для ЭВМ № 2002610852. Зарегистрировано 30.05.2002 г.
7. Попов С.В. Способ поиска информации в политематических массивах неструктурированных текстов. Патент РФ на изобретение № 2266560. Приоритет изобретения 28.04.2004 г.